Intelligent Archives in the Context of Knowledge Building Systems

H. (Rama) Ramapriyan*, G. McConaughy*, C. Lynnes*, K. McDonald*,
S. Kempler*, D. Isaac**

* NASA Goddard Space Flight Center, Greenbelt, MD

** Business Performance Systems, Falls Church, VA

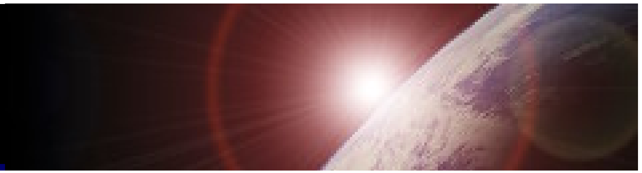
AISRP PI Meeting

April 4-6, 2005

Rama.Ramapriyan@nasa.gov

daac.gsfc.nasa.gov/IDA/

Project Summary



OBJECTIVE

- Design a conceptual architecture for future intelligent data archives that effectively manage and extract knowledge from large volumes of data

APPROACH

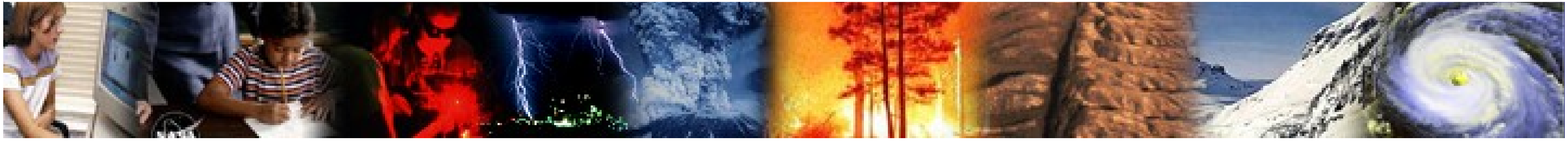
- Collaborate with NASA research projects (IDU/AISRP)
- Derive capabilities & solution concepts from usage scenarios & technology projections
- Validate concepts in operational-scale testbed

PROGRESS

- Identified meaningful usage scenarios
- Identified needed capabilities
- Assessed implementation issues
- Defined functional architecture
- Started testbed implementation

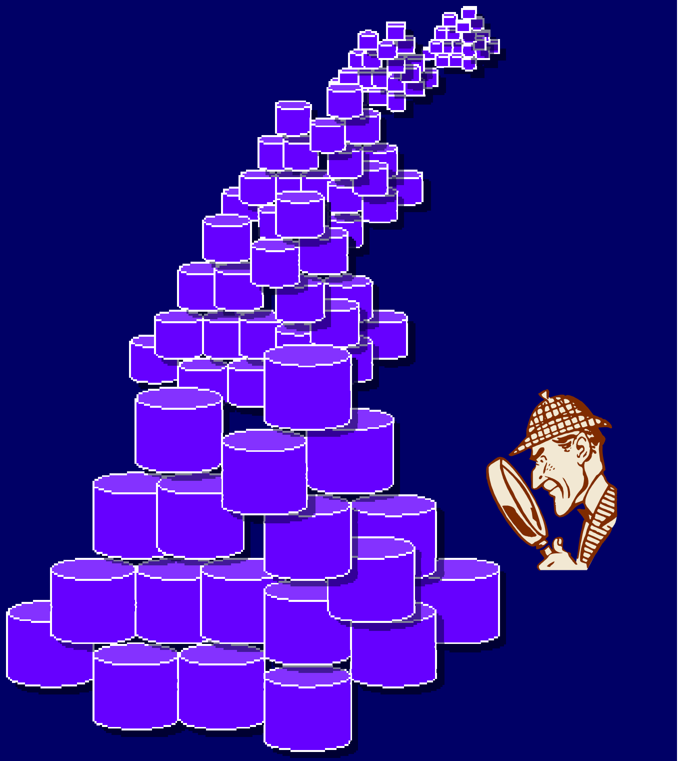
NEXT STEPS

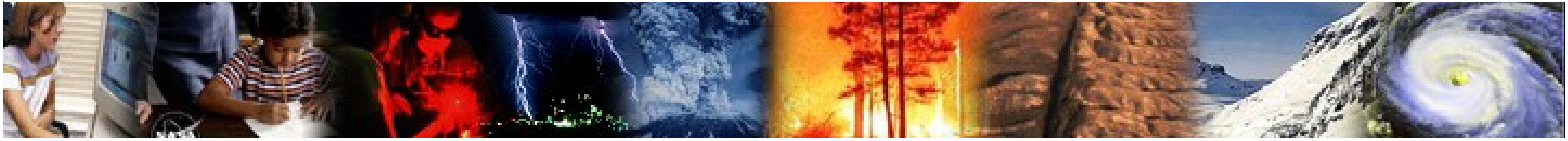
- Implement testbed
- Assess feasibility of large-scale data mining
- Assess science results



Motivation:
Succeeding in a Data Rich Environment

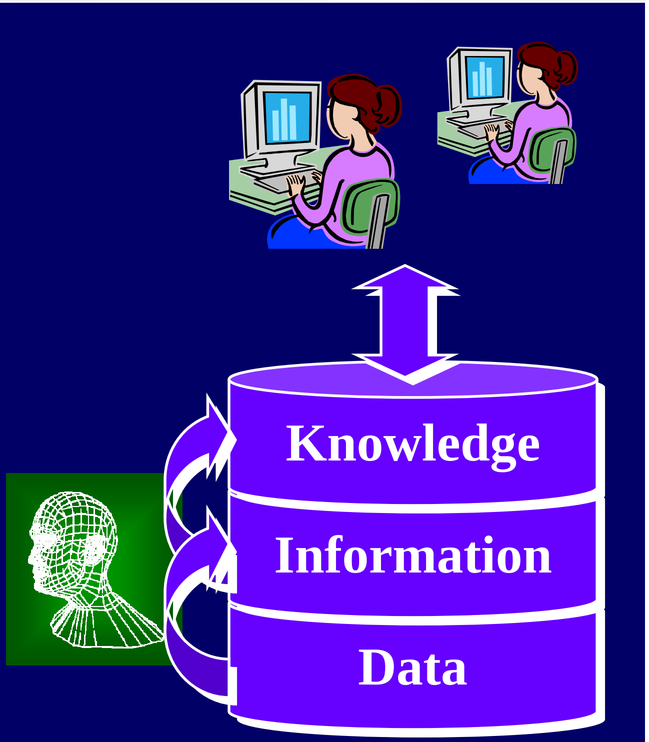
- Large and growing data collections from the Earth Observing System
  - 3.4 petabytes of data
  - 48 million files
  - 3.5 terabytes/day accumulation
- Distributed, heterogeneous data systems
  - 50 data centers
  - Complex value chains
- Broad & diverse user community
  - Research, applications, education
- Limited human capacity to examine large volumes of data
  - Users need information, not just data

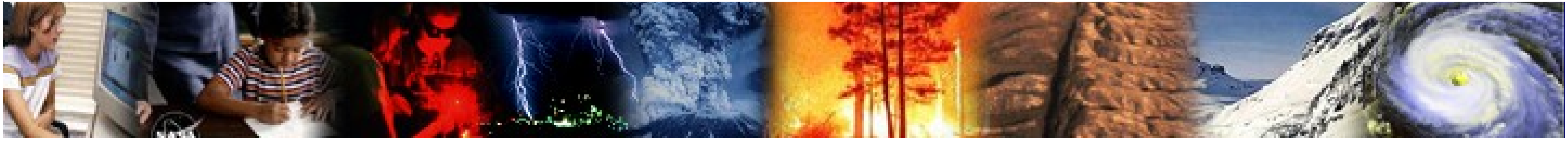




The Concept

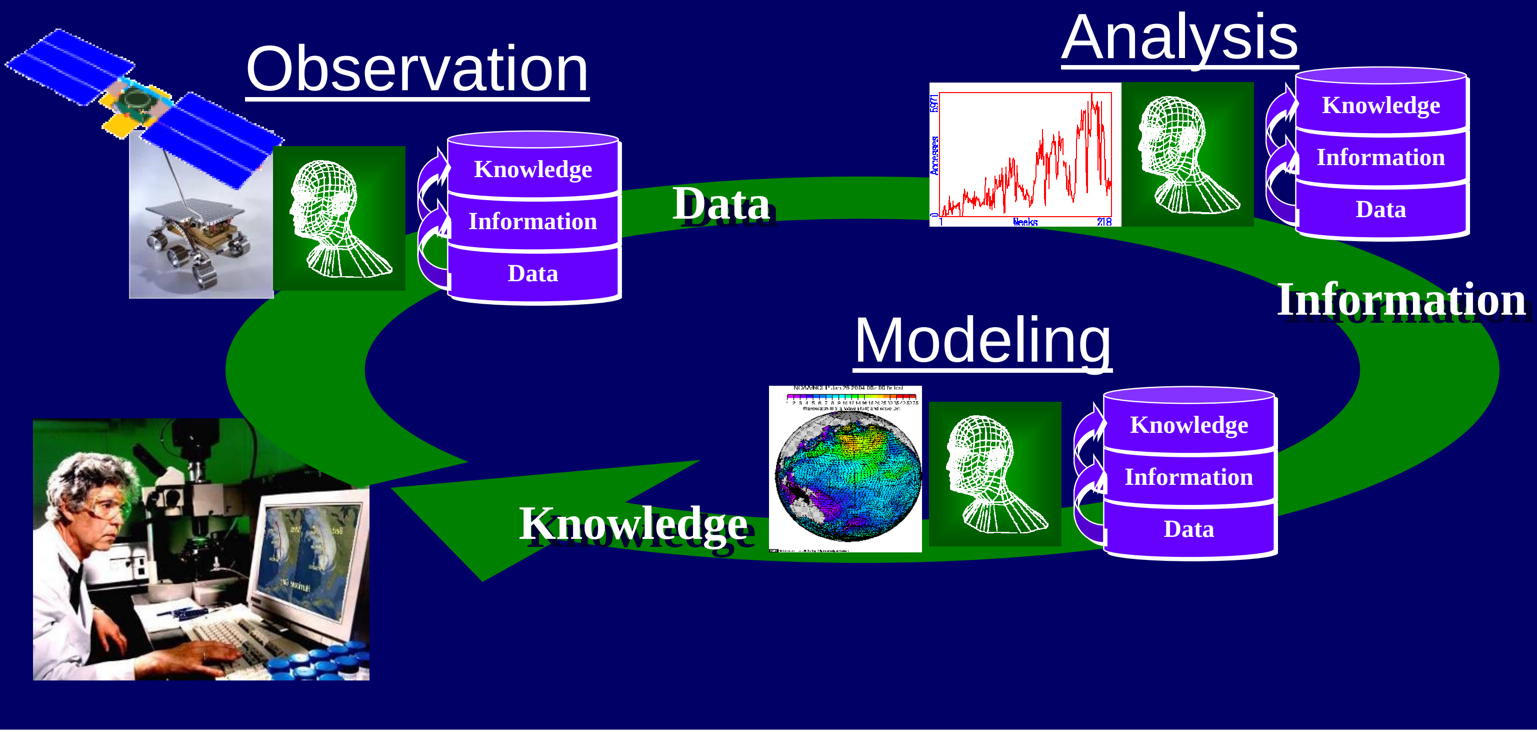
- Intelligent Archives
  - Archive is aware of its own content and usage
  - Archive can extract new information from data holdings
- Knowledge Building Systems
  - Directly support building knowledge from data and information
  - Incorporates intelligent archives to extract information & knowledge
  - Includes feedback loops to improve adaptation to user needs and external events
  - Includes coordination between intelligent archives and intelligent sensors
  - Highly distributed and collaborative





Intelligent Archives in the Context of Knowledge Building Systems (IA-KBS)

- Data archives exist throughout the information value chain
- Intelligence with feedback loops makes systems more effective
- Distributed intelligent components collaborate to achieve user goals



Observation

Knowledge

Information

Data

Data

Analysis

Knowledge

Information

Data

Information

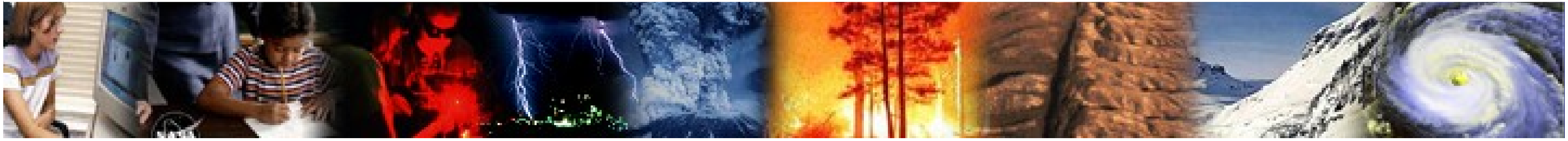
Modeling

Knowledge

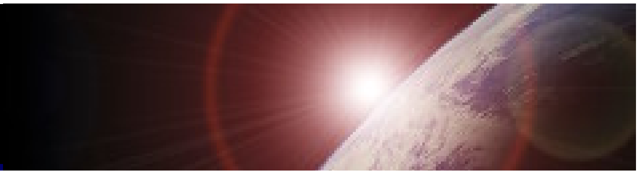
Information

Data

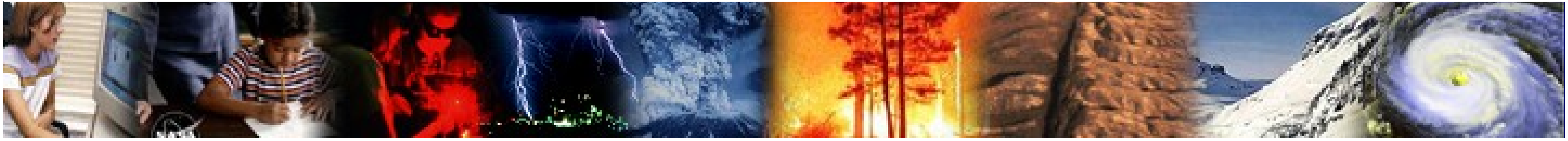
Knowledge



IA-KBS Scenarios

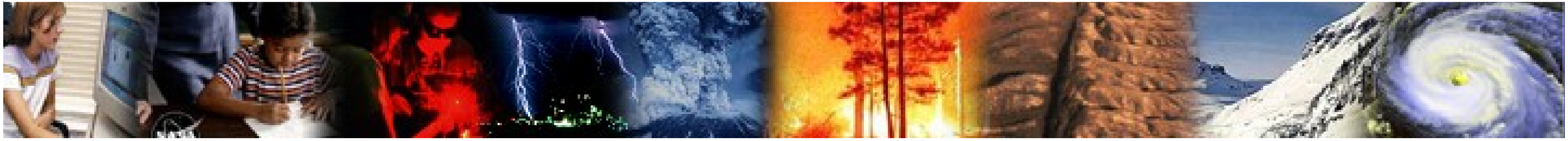


- Advanced weather forecasting
- Precision agriculture
- Virtual observatories
- Wildfire prediction ⟵
- Climate index discovery



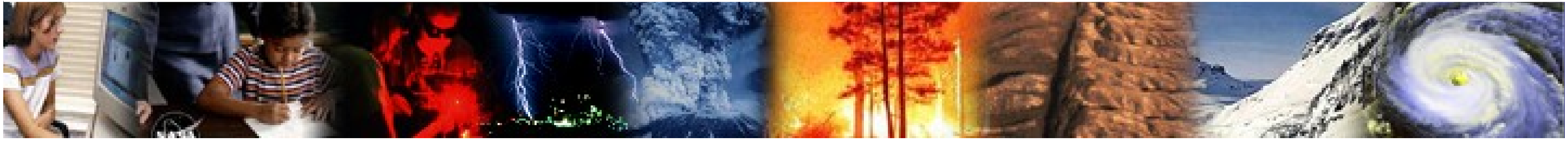
IA-KBS Potential Capabilities

- Virtual product generation
  - Dynamically assemble an information product specific to the user’s need from relevant data
  - Intelligence needed to understand data relationships relative to an information “goal” and anticipate user requests
- Significant event detection
  - Automatically learn “normal” data streams and identify exceptions
  - Intelligent archive can focus attention on interesting data subsets
- Automated data quality assessment
  - Automatically identify anomalies in the data stream
  - Relieves human burden and enables rapid quality assessments

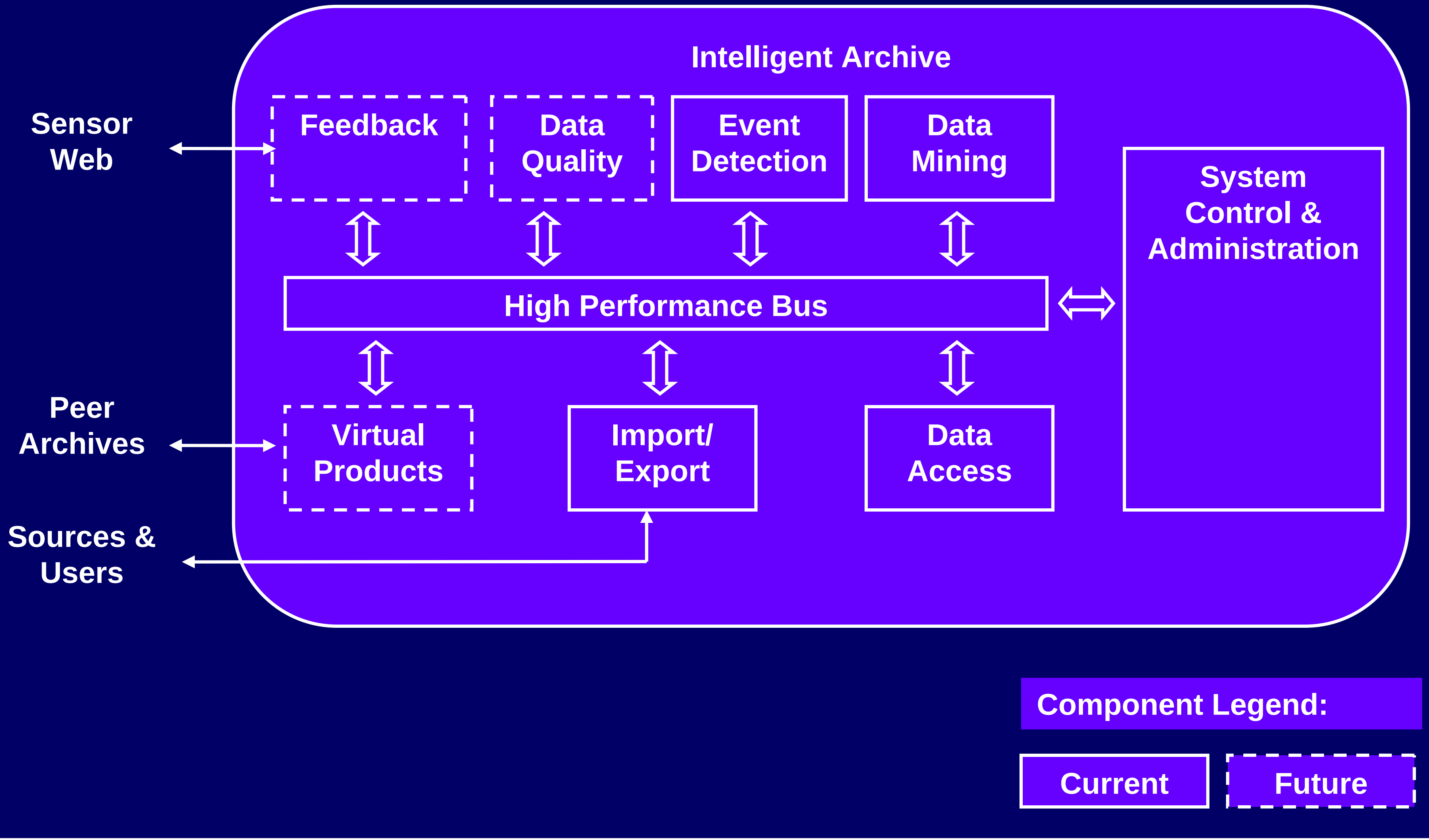


IA-KBS Potential Capabilities (cont'd)

- Large-scale data mining
  - Continuously mine archived data searching for hidden relationships and patterns
  - Enables archive to suggest models for human evaluation
- Dynamic feedback loop
  - Acting on information discovered, such as a significant event
  - Enables archive to adapt to events and anticipate user needs
- Data discovery and efficient requesting
  - Identifying new data sources and information collaborators, and using available resources judiciously
  - Enables archive to reach farther than it's own holdings



Functional Architecture



Intelligent Archive

Feedback

Data Quality

Event Detection

Data Mining

High Performance Bus

Virtual Products

Import/ Export

Data Access

System Control & Administration

Sensor Web

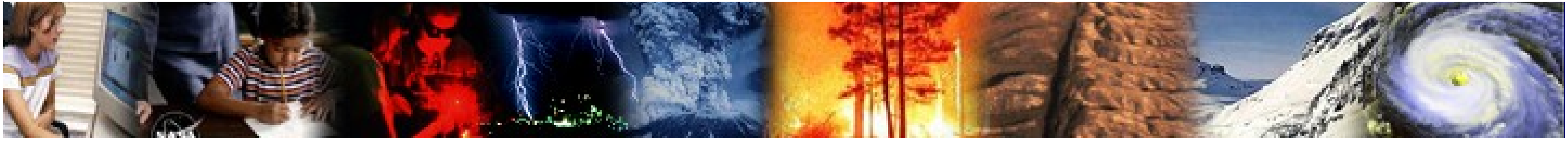
Peer Archives

Sources & Users

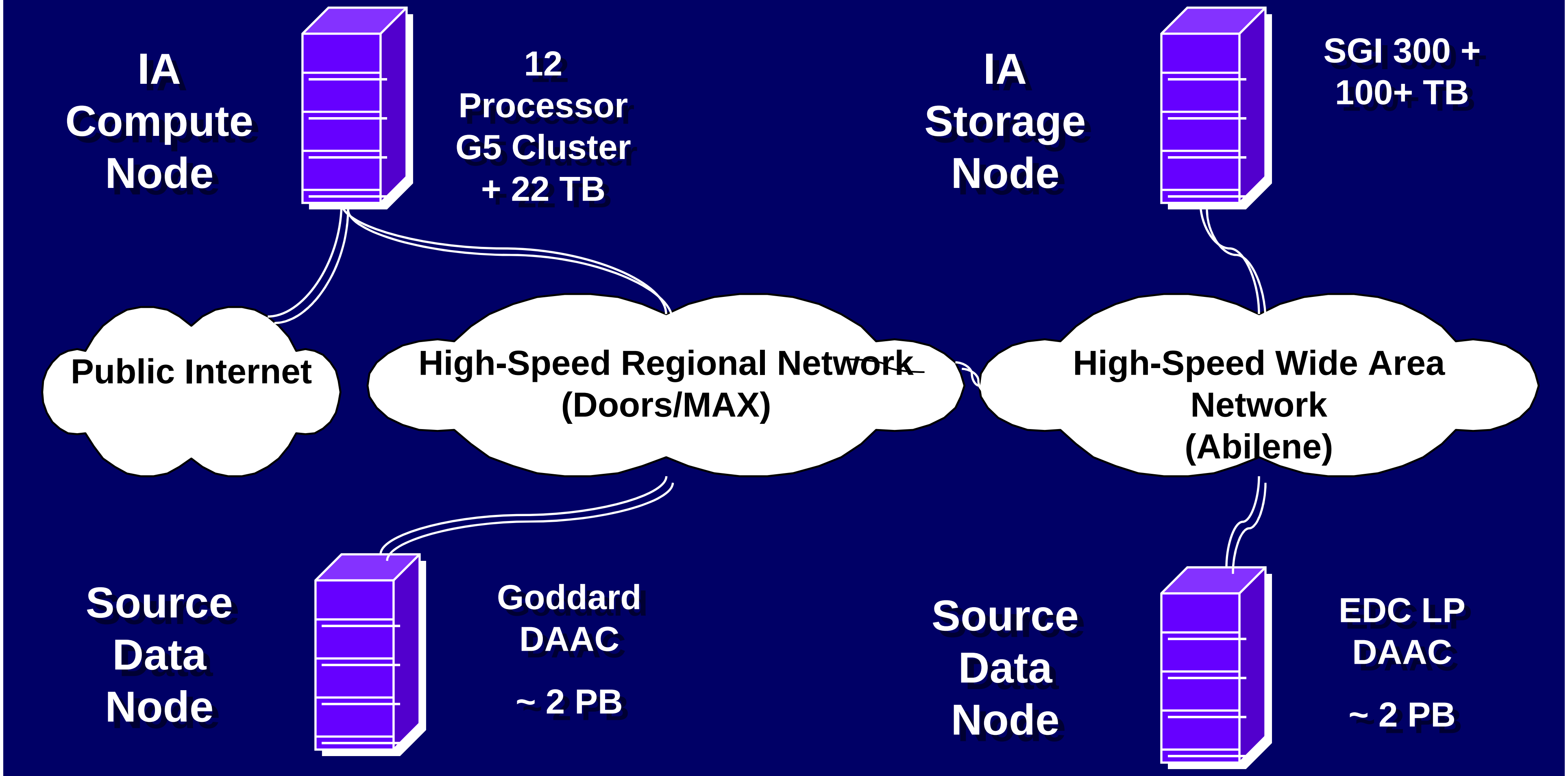
Component Legend:

Current

Future



System Network Architecture



IA Compute Node

12 Processor G5 Cluster + 22 TB

IA Storage Node

SGI 300 + 100+ TB

Public Internet

High-Speed Regional Network (Doors/MAX)

High-Speed Wide Area Network (Abilene)

Source Data Node

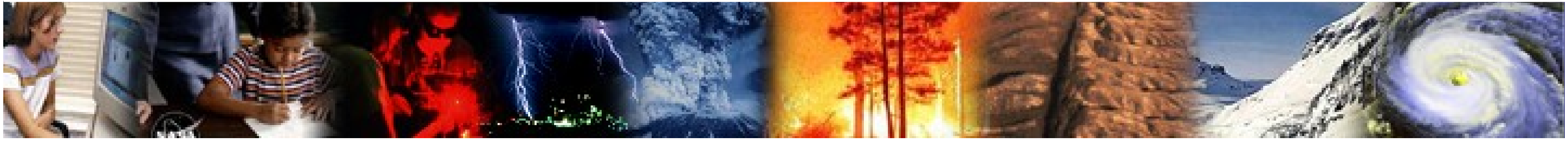
Goddard DAAC

~ 2 PB

Source Data Node

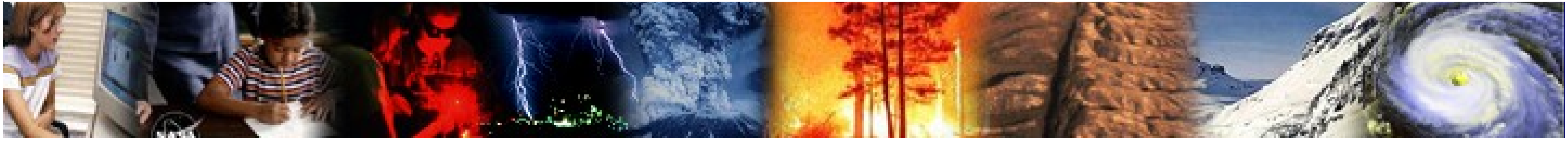
EDC LP DAAC

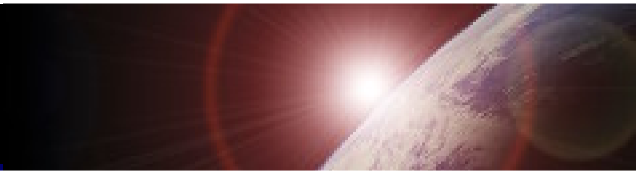
~ 2 PB



IA-KBS – Relevant Technologies

- Distributed system architectures
  - Especially, Grid technologies
- Intelligent data understanding algorithms
  - Fern & Brodley: understanding high-dimensionality data using clustering, re-projection, cluster ensembles
  - Kumar et al: discovering climate indices using clustering on time-series data
  - Danks et al: ecosystem prediction with identification & analysis of extreme events
  - Teng: identifying and removing anomalies to improve classifier performance
  - Kargupta: extending data mining algorithms to distributed architectures
  - Smelyanskiy: Bayesian inference of non-linear dynamical model parameters
  - Nemani & Golden: dynamic assembly of data and operators to satisfy a user’s information goal
  - LeMoigne: sub-pixel accurate image registration for data fusion





Conclusions

- Intelligent archives can improve the utility of data
  - Improved timeliness, ease of access, understandability, readiness for use, and responsiveness
- Intelligent archives can enable a variety of needed capabilities
  - Virtual Product Generation, Significant Event Detection, Automated Data Quality Assessment, Large-Scale Data Mining, Dynamic Feedback Loop, and Data Discovery and Efficient Requesting.
- Promising data mining algorithms have been identified and applied to remote sensing data in a laboratory environment
- Next step is to demonstrate utility and scalability in an operational environment